

STAT 423/623

PROBABILITY IN BIOINFORMATICS AND GENETICS

Katie Kyunghyun Lee

Spring 2019

Contact Information

Instructor: Katie Kyunghyun Lee
Office: DCH 1038
Email: kl25@rice.edu
Office Hours: TBD

Course Objectives and Learning Outcomes

Advances in computing and molecular biology and genomics/genetics, proteomics and so forth have had an immense impact on the biomedical fields, with broad consequences for humanity. Correspondingly, new areas of probability and statistics have been developed specifically to meet the needs of this area. It is now necessary to teach a course that introduces probability and statistics in the modern biology context. Probabilistic models are becoming increasingly important in analyzing the huge amount of data being produced by large-scale DNA-sequencing efforts such as the Human Genome Project or individual peoples genomes. For example, hidden Markov models are used for analyzing biological sequences, linguistic-grammar-based probabilistic models for identifying RNA secondary structure, and evolutionary models for inferring of sequences from different organisms. This course also describes some of the main statistical applications in the field, including BLAST, gene finding, and evolutionary inference. Further topics include microarrays, molecular evolution and stochastic processes for genomics.

This course is offered by the department of Statistics. Therefore, only a basic understanding of biological concepts is assumed, and all concepts are explained when used or can be understood from the context.

Students will be able to

1. have a strong understanding of fundamental concepts of the statistics that underlie bioinformatics;
2. understand the formulation and proof of stochastic models for genomic data;

3. apply statistical techniques (commonly employed in bioinformatics) to analyze microarray data and interpret the results generated;
4. use statistical software (R, SPSS, etc) for solving complex problems in bioinformatics;
5. interpret the statistical results as reported in the bioinformatics literature.

Required Texts and Materials

- Bioinformatics, by Polanski and Kimmel (Springer 2007). Available electronically from Fondren.
- Ewens and Grant (2005) Statistical methods in bioinformatics (2nd Ed.). Springer

Grade Policies

Assignment:

- 8 to 10 homeworks (40%): Mostly from the *Problem* chapter in the textbooks
- All assignments will be graded within one week, or at least no more than 10 days after the due date. Some students might turn in the assignment earlier than the due date. I usually grade all assignment together unless you make specific requirement.
- Homework can be discussed in groups. If you receive help from other students, then specify their names and briefly explain how you collaborate. Final redaction has to be done personally. Verbatim copying of homework assignments is a violation.

Exam:

- Two in-class tests: midterm (20%) and cumulative final (40%). There is no make-up exam. Each exam time limit is approximately 2 hours.
- The difficulty of the tests will be very similar to that of homeworks. If you do your best with homeworks, then you will be able to get good grades in the tests.
- Tests are under strict Honor Code rules and have to be completed without collaboration or help from others.

Other information:

- Only for the Stat 623 students: Final project with a presentation (P/F)
- NO extra credits will be awarded for individual students. Please do not attempt to ask for individual extra credit at the end of semester to raise your grade. It is not fair to other students.

Miscellaneous

- Attendance: Essential, though not enforced administratively
- There will be weekly (or biweekly) TA sessions.
- You can just call me Katie since I am not a PhD *yet* but a student just like you :-)
- This is the *very first* class that I am teaching as an instructor. Therefore, your feedback and suggestion are very important to me, and I will make sure that everyone's voice is not missed!

Rice Honor Code

In this course, all students will be held to the standards of the Rice Honor Code, a code that you pledged to honor when you matriculated at this institution. If you are unfamiliar with the details of this code and how it is administered, you should consult the Honor System Handbook at <http://honor.rice.edu/honor-system-handbook/>. This handbook outlines the University's expectations for the integrity of your academic work, the procedures for resolving alleged violations of those expectations, and the rights and responsibilities of students and faculty members throughout the process.

Disability Resource Center

If you have a documented disability or other condition that may affect academic performance you should: 1) make sure this documentation is on file with the Disability Resource Center (Allen Center, Room 111 / adarice@rice.edu / x5841) to determine the accommodations you need; and 2) talk with me to discuss your accommodation needs.

Syllabus Change Policy

This syllabus is only a guide for the course. Circumstances and events, such as student progress, may make it necessary for the instructor to modify the syllabus during the semester. Therefore, this syllabus is subject to change with advanced notice.

Course Schedule

Outline :

- Single DNA sequence: Shotgun sequencing, Markov-Chain modeling, patterns and repeats.

- Multiple DNA or protein sequences: Dynamic programming alignments, substitution matrices and relationships with evolution of structure and function of proteins.
- Evolutionary Models: Models of nucleotide mutation and substitution, recombination and genetic drift, with applications to genome evolution and population genetics.
- Stochastic models of evolution and progression of genetic disease.
- Molecular phylogenetics (tree making): distance matrix, maximum likelihood and parsimony.
- BLAST: Algorithm for sequence retrieval and comparison.
- Hidden Markov Models: Viterbi Algorithm and its applications.
- Systems biology: Analysis and modeling of gene and protein networks.

January 8 & January 10

Introduction (Basic concepts in biology)

January 15

Shotgun sequencing and coverage

January 17 & January 22 & January 24

Elementary introduction to Markov chains

January 29 & January 31

Molecular evolution

February 5

Analysis of pairs of sequences (DNA)

February 7

No class (Spring recess)

February 12

(Continued) Analysis of pairs of sequences (DNA)

February 14 & February 19

Modeling signals in DNA

February 21 & February 26 & February 28

Population genetics

March 5

Problem solving session

March 7

Midterm exam day!

March 12 & 14

No class (Spring break)

March 19 & March 21 & March 26

Phylogenetics

March 28

RNA folding problem

April 2 & April 4

Random Walk and BLAST

April 9 & April 11 & April 16

Viterbi Algorithm and Hidden Markov Chain

April 18

Final presentation day for STAT 623

April 25

Final exam day!